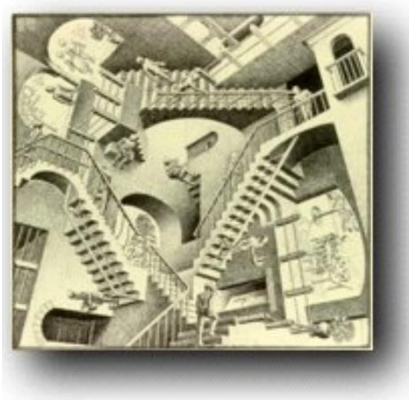
Example: Financial Stability *(example from Johnson and Wichern, p.655 – data is bankrupt.csv). Four ratios of financial health were measured for 46 companies:*

$$\frac{\text{Cash Flow}}{\text{Total Debt}} \quad \frac{\text{Net Income}}{\text{Total Assets}} \quad \frac{\text{Current Assets}}{\text{Current Liabilities}} \quad \frac{\text{Current Assets}}{\text{Net Sales}}$$

The data measures data for the prior two years for 25 companies that went bankrupt and 21 companies that were financially sound.

Disciminant Analysis in SAS. Use `PROC DISCRIM.`
I've put code on CANVAS to produce the output we'll discuss in class.

Stepwise DA



- So far we've assumed that **all variables** considered are good discriminators of groups.

- If we have many variables, we would like to identify **which variables actually discriminate well between groups** (i.e. which variables should be used in the discriminant function)

- Three solutions: **forward, backward, stepwise selection procedures** (like stepwise or best subsets regression)

Forward	Backward	Stepwise
• Start with no variables • Add the one variable that provides 'best' discrimination • Keep adding variables until addition of more variables does not improve the discrimination of the model	• Start with all variables • Remove the one variable that 'least' reduces the discrimination • Keep removing variables until removal of more variables causes 'unacceptable' reduction in discrimination	• Start with no variables • At each stage, either add a 'best' variable if addition of a variable causes 'significant' improvement in discrimination, or remove a variable if removal does not cause a 'significant' reduction in discrimination • Continue until no variables are added or removed

The stepwise selection method is generally considered the best method

Example: Land Cover. *Land cover is classified into two functional groups: forbs (1) and shrubs (2). We have five possible discriminators: are these all actually necessary and helpful?*

- *N - concentration of nitrogen*
- *AMASS - mass-based net photosynthetic capacity*
- *AAREA- area-based net photosynthetic capacity*
- *GS - leaf diffuse conductance at photosynthetic capacity*
- *LSLA - log10 transformation of specific leaf area*

Stepwise Selection Criteria

Wilks' Lamdba: recall that this is defined as

$$\Lambda = \frac{|\mathbf{SSCP}_w|}{|\mathbf{SSCP}_t|}$$

At each step add/remove variables according to whether they make Wilks' Lamdba 'meaningfully' smaller/larger (remember that a smaller value indicates greater separation between groups).

Rao's V – based on Mahalanobis distance, tries to maximize MD distance of centroids from overall centroid. Again, add/remove variables according to a threshold change.

Between Groups F-ratio – again, uses MD distance, but accounts for differences in sample sizes between groups. Again, add/remove variables according to a threshold change.

Wilks' Lamdba is the most common selection criteria

For **stepwise** DA, need to establish **threshold values for entry, removal of model terms:**

- At each step, add discriminating variable if model improves above a set threshold.
- At each step, remove a discriminating variable if model criteria does not move below a set value
- Continue till no variables are added/removed according to cut-off values.

- **For Wilks' Lambda, cut-off values can be converted to threshold F-values (i.e. p-values)**

Computer Notes: Stepwise DA. In SPSS, click on Stepwise. Under Method, select selection criteria method. In SAS, use separate procedure STEPDISC. Based on results of STEPDISC, run DISCRIM or CANDISC. STEPDISC only uses Wilks' Lambda selection criteria. In R, you can use the `stepclass()` function in the `klaR` package (example on CANVAS).

Example: Land Cover. *Start by looking that the correlation of the possible discriminating variables:*

	N	AMASS	AAREA	GS
AMASS	0.851			
AAREA	0.111	0.443		
GS	-0.255	-0.109	0.551	
LSLA	0.919	0.822	-0.035	-0.240

Sample correlations reveal that several variables are highly correlated. This suggests that not all of the variables will necessarily be good discriminators.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
N	.710	8.981	1	22	.007
AMASS	.749	7.355	1	22	.013
AAREA	.949	1.175	1	22	.290
GS	.965	.802	1	22	.380
LSLA	.789	5.893	1	22	.024

At the single variable level, not all are good discriminators.

Stepwise Statistics

Variables Entered/Removed[a,b,c,d]

Step	Entered	Wilks' Lambda							
						Exact F			
		Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	N	.710	1	1	22.000	8.981	1	22.000	.007

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 10.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

Variables in the Analysis

Step		Tolerance	F to Remove
1	N	1.000	8.981

Only Nitrogen is selected as a good discriminator. After this variable, none of the remaining variables makes a significant improvement in discriminating ability (F-statistics are all below the threshold F-value of 3.84)

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	N	1.000	1.000	8.981	.710
	AMASS	1.000	1.000	7.355	.749
	AAREA	1.000	1.000	1.175	.949
	GS	1.000	1.000	.802	.965
	LSLA	1.000	1.000	5.893	.789
1	AMASS	.397	.397	.253	.702
	AAREA	.994	.994	.501	.694
	GS	.920	.920	2.241	.642
	LSLA	.284	.284	.028	.709

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.710	7.360	1	.007

Multivariate means are not the same across groups.

Of course at this point, this is basically a t-test of nitrogen between two groups!

Classification Results[a]

			Predicted Group Membership		
		FUNCTION	1	2	Total
Original	Count	1	9	5	14
		2	0	10	10
	%	1	64.3	35.7	100.0
		2	.0	100.0	100.0

a. 79.2% of original grouped cases correctly classified.

Stepwise prediction is just as good as with all variables (79%). However, now we only use one discriminating variable. This is not always the case – sometimes, classification accuracy will decrease. The problem is analogous to that of R-squared: if you include more predictors, the R-squared will increase, even if predictors are not significant.

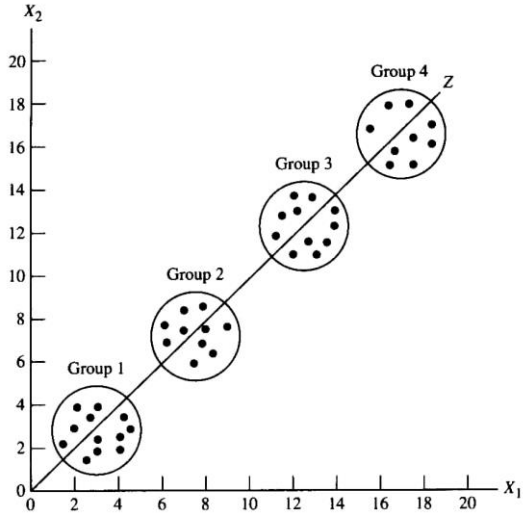
Go to R to see example of stepwise quadratic DA.

Multiple (3+) Group Discriminant Analysis (MDA)

Still have two goals:

- **Discrimination** – identify function(s) that are linear combinations of observed variables that discriminate between groups
- **Classification** – for new observations, use results of discrimination (*or not – also a separate procedure*) to assign group membership

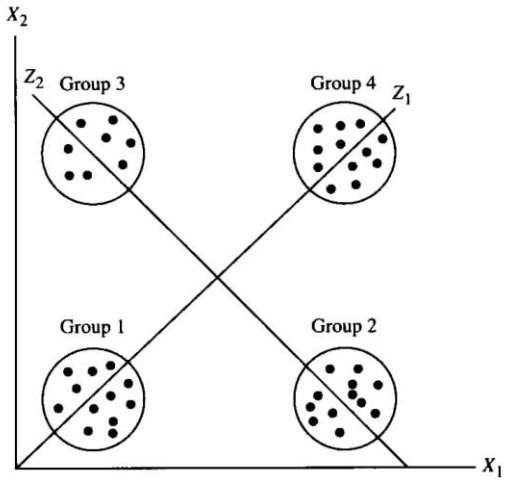
An illustrative picture: consider four groups and two variables.



Case 1: one discriminant function is sufficient – projection onto Z provides reasonable separation (for classification, choose three cut points)

Case 2: first discriminant function Z_1 provides good separation between groups 1 and 4, but not between groups 2 and 3 – a second discriminant function Z_2 is required to distinguish between groups 2 and 3.

Goal: discriminate between groups with relatively few discriminant functions.



Determining Discriminant Functions

Can read more HERE and also Johnson and Wichern, 11.7

Goal: find weights to maximize the sum of squares between groups relative to the sum of squares within groups (use sum of squares and cross products matrices):

$$\lambda = \frac{\mathbf{w}'\mathbf{SSCP}_b\mathbf{w}}{\mathbf{w}'\mathbf{SSCP}_w\mathbf{w}}$$

Taking partial derivatives with respect to $\mathbf{w}$ and setting to zero, we have a maximum when

$$\left(\mathbf{SSCP}_w^{-1}\mathbf{SSCP}_b - \lambda\mathbf{I}\right)\mathbf{w} = 0$$

an **eigenvalue** problem with associated **eigenvectors w.**

Assuming that no variable is a linear combination of other variables, the **number of eigenvalue/eigenvector solutions** to the non-symmetric matrix $\mathbf{SSCP}_w^{-1}\mathbf{SSCP}_b$ is

$$\min((G-1), p)$$

where G is the number of groups and p is the number of variables.

- Resulting discriminant functions are not necessary uncorrelated (i.e. orthogonal).
- However, discriminant scores **are** uncorrelated.

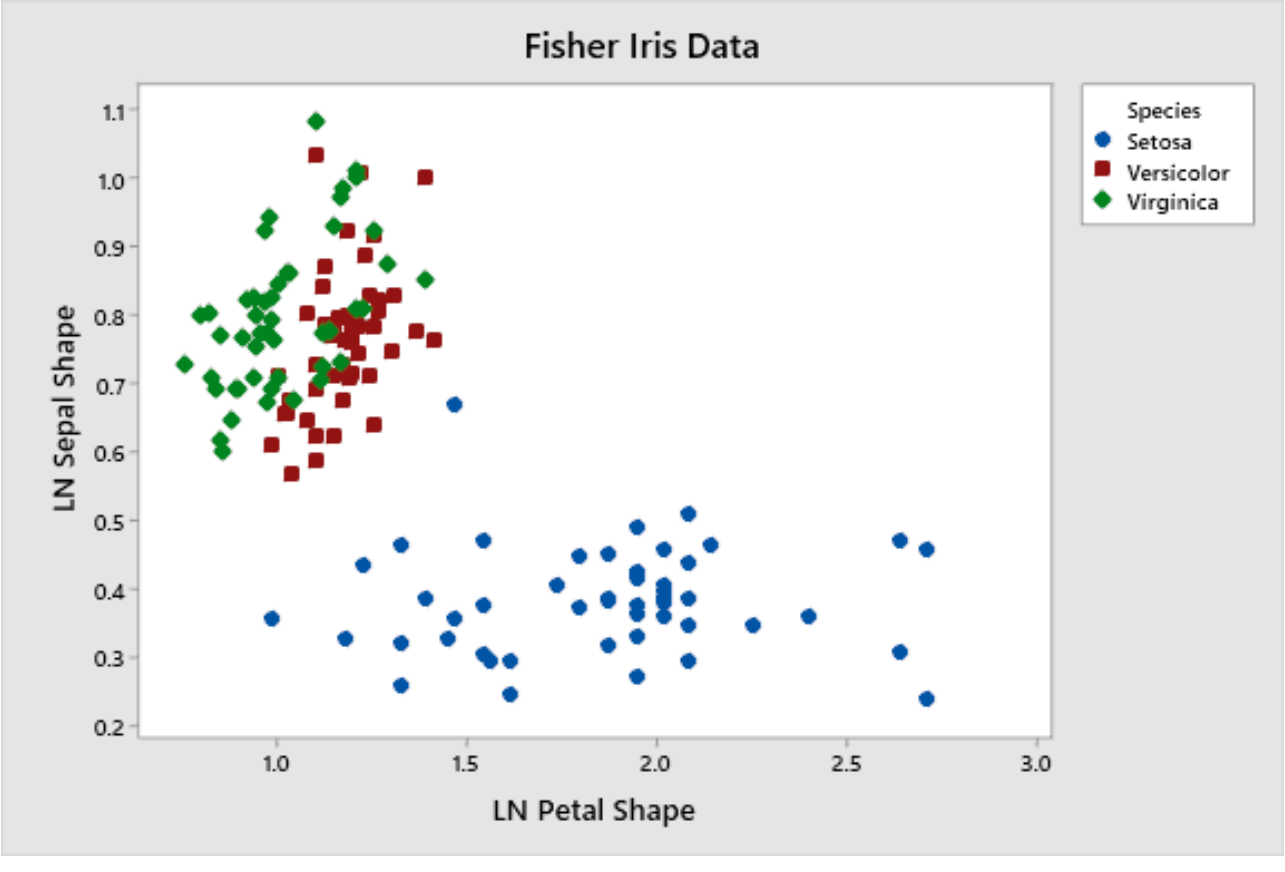
Eigenvectors are ordered according to **discriminating ability**:

- The **first** eigenvector/eigenvalue gives the discriminant function of **maximum separation** among groups.
- Subsequent eigenvectors/eigenvalues give the discriminant function of maximum separation among groups such that the discriminant scores for this function are uncorrelated with all previous eigenvector scores.
- The eigenvalues give the ratio of $\frac{\mathbf{w}'\mathbf{SSCP}_b\mathbf{w}}{\mathbf{w}'\mathbf{SSCP}_w\mathbf{w}}$ for each linear combination.

Computer Notes: MDA. Basically the same as two group DA. Group variable has three or more levels instead of 2 levels.

Example: Fisher Iris Data *(famous example in discriminant analysis, cluster analysis). Data published (1936) by R.A. Fisher, parent of modern statistics. The log(ratio) of sepal length/sepal width and log(ratio) of petal length/petal width were measured in millimeters on fifty iris specimens from each of three species:* Iris setosa, I. versicolor, and I. virginica.

A plot of the species:

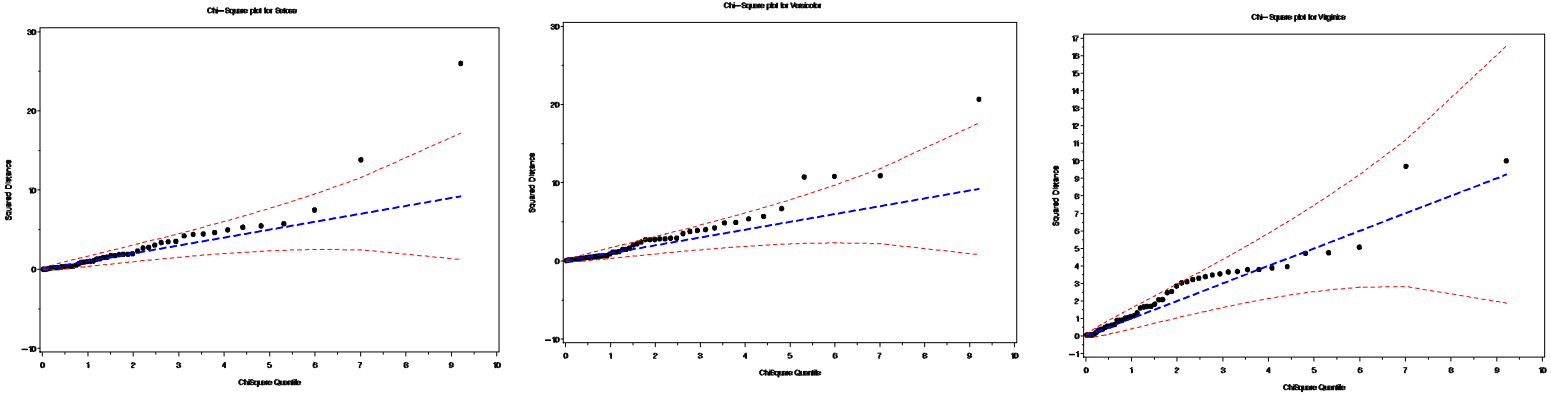


Goal – use discriminant analysis to identify axes of differences among groups.

MDA Assumptions: same as two group MDA

- Each group has a multivariate normal distribution
- Each group has an identical covariance matrix (if not, use quadratic discrimination)

Example: *Iris Data. Make Chi-square plot for sum of square variables for each group. Plots seem reasonably linear.*



HOWEVER – plot of data suggests that Setosa has a somewhat different covariance matrix than Virginica and Versicolor.

Box's Test of Equality of Covariance Matrices

Log Determinants

VAR00007	Rank	Log Determinant
1.00	2	-6.995
2.00	2	-9.494
3.00	2	-8.719
Pooled within-groups	2	-7.523

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Test Results

Box's M		129.333
F	Approx.	21.132
	df1	6
	df2	538562.8
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

*This suggests that covariance matrices are NOT equal – might want to use **quadratic** discrimination functions.*

How Many Discriminant Functions?

To answer this, we review how to test equality of means:

Univariate Tests: for each variable test

$$H_0 : \mu_1 = \mu_2 = ... = \mu_G$$
$$H_a : \text{at least one } \mu_i \text{ different from others}$$

In ANOVA, we'd do an overall F-test of means. Here, use **Wilks' Lambda**, approximated by an F-test

Multivariate Test: test **mean vectors**

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{...} = \boldsymbol{\mu}_G$$
$$H_a : \text{at least one } \boldsymbol{\mu}_i \text{ different from others}$$

Use the Multivariate Version of **Wilks' Lambda**:

$$\Lambda = \frac{|\mathbf{SSCP}_w|}{|\mathbf{SSCP}_t|} \text{ where } |\ | \text{ is the determinant}$$

This can be approximated by a chi-square statistic:

$$\chi^2_{p(G-1)} = -\left[n - 1 - \left(\frac{p+G}{2}\right)\right] \ln \Lambda$$

Re-write as

$$\chi^2 = -\left[n - 1 - \left(\frac{p+G}{2}\right)\right] \sum_{k=1}^{K} \ln(1+\lambda_k)$$

where K is the total number of discriminant functions (i.e. $K = \min(p, G-1)$), and λ_k is the eigenvalue of the kth discriminant function.

This is a simultaneous test of the significance of all K discriminant functions

***H*$_o$:** None of the K discriminant functions is a significant discriminator (i.e. all means are equal)

***H*$_a$:** At least the first discriminant function is a significant discriminator (since functions are ordered according to maximum separation).

NOW: if the first function is significant, test functions 2 through K:

$$\chi^2 = -\left[n - 1 - \left(\frac{p+G}{2}\right)\right]\sum_{k=2}^{K}\ln(1+\lambda_k)$$

***H*$_o$:** None of the discriminant functions 2 through K is a significant discriminator

***H*$_a$**: At least the second discriminant function is a significant discriminator

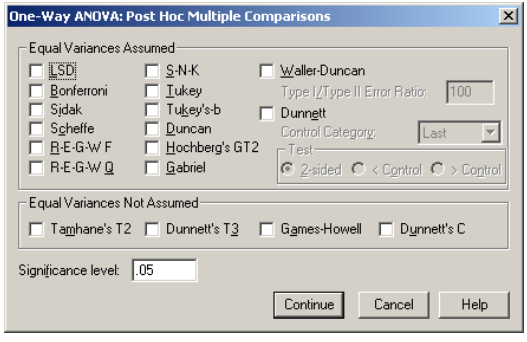
Test for discriminant functions $r \leq K$ through K:

$$\chi^2 = -\left[n - 1 - \left(\frac{p+G}{2}\right)\right]\sum_{k=r}^{K}\ln(1+\lambda_r)$$

Aside:

- For each discriminant function, we are getting a test of whether or not all groups have the same mean discriminant function scores (think ANOVA – null hypothesis is that all means are equal, alternative hypothesis is that they are not all equal).
- Might like to know which group means are not equal
- Solution:
    - Obtain discriminant scores for each function and SAVE.

- Use univariate multiple comparison procedures for each each functions' scores to evaluate differences in pairs of means (i.e. Tukey or Scheffe' comparisons)
- In SPSS, use Analyze → Compare Means → One Way ANOVA, choose Post Hoc Multiple Comparisons . In SAS, use PROC GLM and use MEANS statement. R example code shows how to calculate scores.



***Example**: Iris data. The univariate ANOVA type test of means for each variable:*

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
LOGY1	.205	284.312	2	147	.000
LOGY2	.305	167.761	2	147	.000

Successive tests for groups of discriminant functions:

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.115	316.548	4	.000
2	.986	2.135	1	.144

This suggest that only the first discriminant function is really a significant discriminator – i.e., we're having trouble finding any function that discriminates well between virginica and versicolor.

NOW - a bit of math

We measure responses for G groups on p variables. The number of discriminant functions is

$$\min((G-1), p)$$

These discriminant functions ('mostly orthogonal' directions of maximum discrimination) have corresponding eigenvalues from

$$\mathbf{SSCP}_b \mathbf{SSCP}_w^{-1} = \frac{\mathbf{SSCP}_b}{\mathbf{SSCP}_w}$$

That is: each eigenvalue measures Sum of Squares between groups divided by Sum of Squares within groups **in the direction of the *i*th discriminant function:**

$$\lambda_i = \frac{SS_{between\_groups}}{SS_{within\_groups}} = \frac{SSB_i}{SSW_i}$$

In other words: λ_i measures how strongly scores on that function are related to group membership.

NOW: Let's define a new quantity – the **CANONICAL CORRELATION:**

$$\rho_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}} \quad \text{or} \quad \lambda_i = \frac{\rho_i^2}{1-\rho_i^2}$$

A little algebra:

$$\rho_i^2 = \frac{\lambda_i}{1+\lambda_i} = \frac{\frac{SSB_i}{SSW_i}}{1+\frac{SSB_i}{SSW_i}} = \frac{\frac{SSB_i}{SSW_i}}{\frac{SSW_i}{SSW_i}+\frac{SSB_i}{SSW_i}}$$

$$= \frac{SSB_i}{SSW_i + SSB_i} = \frac{SSB_i}{SST_i}$$

The squared canonical correlation ρ_i^2 measures the amount of variability in the direction of *i*th discriminant function that is explained by group membership!

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	7.552[a]	99.8	99.8	.940
2	.015[a]	.2	100.0	.120

a. First 2 canonical discriminant functions were used in the analysis.

For each discriminant function, compute the canonical correlation of that function from the associated eigenvalue:

$$\rho_k = \sqrt{\frac{\lambda_k}{1+\lambda_k}}$$

Can also use eigenvalues to calculate **percent of variation**: relative importance of each discriminant function

$$\frac{\lambda_k}{\sum_{k=1}^{K} \lambda_k}$$

For Iris data, percent variation for each discriminant function is 99.8% and .2%, respectively.

Coefficients

Like two group DA, we have unstandardized, standardized, and (if calculated by you), normalized.

Iris Data: Standardized coefficients suggest that both ratios are of relatively equal importance

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
LOGY1	.852	.559
LOGY2	-.712	.729

Canonical Discriminant Function Coefficients

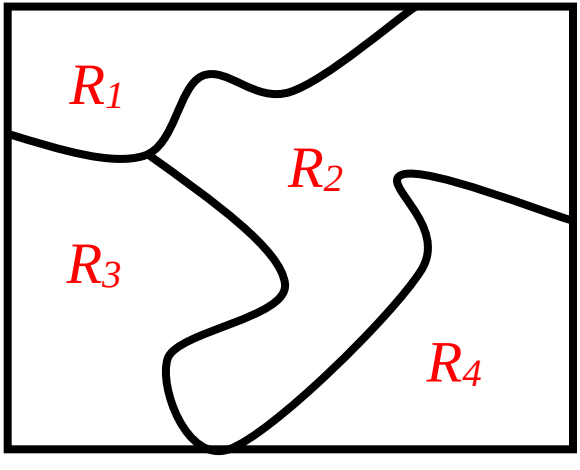
	Function	
	1	2
LOGY1	8.813	5.781
LOGY2	-2.907	2.974
(Constant)	-1.793	-7.752

Unstandardized coefficients

Classification

Extension of Decision Theory to three or more groups

- Let Ω be the p-dimensional sample space of all possible observations $\mathbf{x}$
- Let $f_i(\mathbf{x})$ be the probability density function associated with group i defined over Ω
- Partition Ω into regions R_1, R_2, … R_G
- For observations $\mathbf{x}$ in R_i, classify an observation as in group i



Misclassification probabilities: Conditional probability of misclassifying an observation from group i in group k:

$$P(k \mid i) = \int_{R_k} f_i(\mathbf{x})\, d\mathbf{x}$$

Prior probabilities: if unequal, usually based on sample size. Denoted by p_i, such that $\sum_{i=1}^{G} p_i = 1$

Misclassification Costs

$c(k \mid i)$ = cost of misclassifying an observation from group i in group k:

Computer Notes: Specifying prior probabilities/costs. In MINITAB, click on OPTIONS. In SPSS, click on `CLASSIFY` (can only base on group sizes). In SAS, use the `PRIORS` statement in `PROC DISCRIM`. In R, use `priors=c(….)` in the `lda()` function. Just like 2-group DA!

Expected Costs of Misclassification (ECM):

Calculate for each group: Cost of misclassifying observation from group 1 into any other group:

$$ECM_1 = c(2|1)P(2|1) + c(3|1)P(3|1) + .. + c(G|1)P(G|1)$$
$$= \sum_{k=2}^{G} c(k|1)P(k|1)$$

Overall ECM : weighted by prior probabilities for each group:

$$ECM = p_1 ECM_1 + p_2 ECM_2 + \ldots + p_G ECM_G$$

$$= \sum_{i=1}^{G} p_i \left(\sum_{\substack{k=1 \\ k \neq i}}^{G} c(k \mid i) P(k \mid i) \right)$$

Goal: pick regions R_i to minimize Overall ECM.

Fact: To minimize ECM, assign each point x to the region where $\sum_{\substack{i=1 \\ i \neq k}}^{G} p_i f_i(\mathbf{x}) c(k \mid i)$ is a minimum.

If misclassification costs and prior distributions are equal, **this reduces to**

Fact: To minimize ECM, assign each point x to the region where $f_i(\mathbf{x})$ is a maximum **(i.e. assign to most probable group!)**

Classification for Normal Populations

Use formulas above, substitute sample means/covariances (equal or unequal), get classification rules. Have a quadratic rule for unequal covariance matrices *(see Johnson and Wichern, p. 616-618 for formulas if you like.)*

Classification functions

Same as for two populations - Coefficients have no particular meaning: calculated value in each function for each observation. Allocate observation to group whose classification function is highest

Example: Iris Data. *Classification results for linear and quadratic functions (recall data seems multivariate normal, but covariance matrices appear unequal).*

Equal Covariance Matrices

Classification Results[a]

			Predicted Group Membership			
		VAR00007	1.00	2.00	3.00	Total
Original	Count	1.00	49	1	0	50
		2.00	0	41	9	50
		3.00	0	16	34	50
	%	1.00	98.0	2.0	.0	100.0
		2.00	.0	82.0	18.0	100.0
		3.00	.0	32.0	68.0	100.0

a. 82.7% of original grouped cases correctly classified.

Unequal Covariance Matrices

Classification Results[a]

			Predicted Group Membership			
		VAR00007	1.00	2.00	3.00	Total
Original	Count	1.00	49	1	0	50
		2.00	0	42	8	50
		3.00	0	11	39	50
	%	1.00	98.0	2.0	.0	100.0
		2.00	.0	84.0	16.0	100.0
		3.00	.0	22.0	78.0	100.0

a. 86.7% of original grouped cases correctly classified.

Observations:

- Overall, classification is slightly better using the quadratic discriminant functions: total % correct classification is higher.
- Classification for Group 1 (Setosa) is almost perfect, as expected
- Notice for linear discrimination the inequality in misclassification rates for groups 2 and 3 (18% and 32%). This is typical when applying linear discrimination to covariance matrices where variability is unequal between groups.

Example: Heavy Metals. *The log(concentration) of eight heavy metals was measured in soil samples from three sites. Four samples were taken from each site. The goal is to use discriminant analysis to see which variables discriminate among the sites.* *(data described in Quinn & Keough, Experimental Design and Analysis for Biologists)*

(worked out in class, data and description is online for you to play with).

Discriminant Score Plots

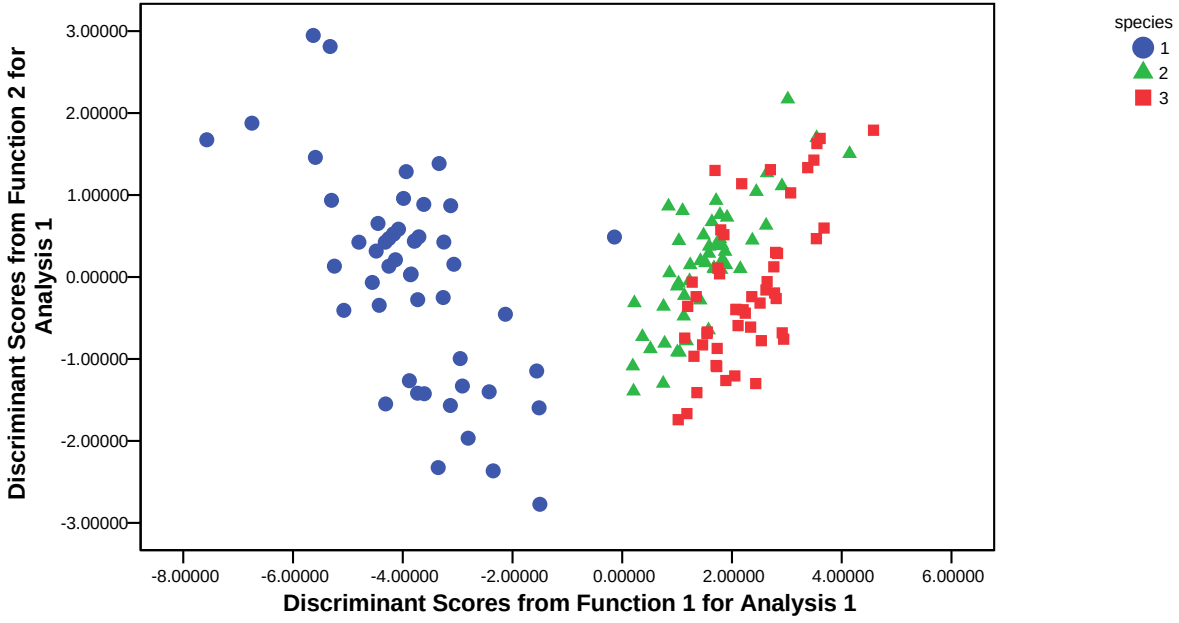
- If **LINEAR** discriminant analysis results in 2 - 4 significant discriminant functions, you can plot observations in **'linear discriminant space'** – i.e. project the observations onto the discriminant axes and plot. This gives you a visual idea of how well the discrimination will work. *If you only have one discriminant function, this is not very interesting.*

- Process –save the linear **discriminant function scores** (these are the projections of data points onto the discriminant axes). Plot Score 1 on the first axis and Score 2 on the second axis (if more scores, make more bivariate plots, i.e. Score 1 vs. Score 3, etc).

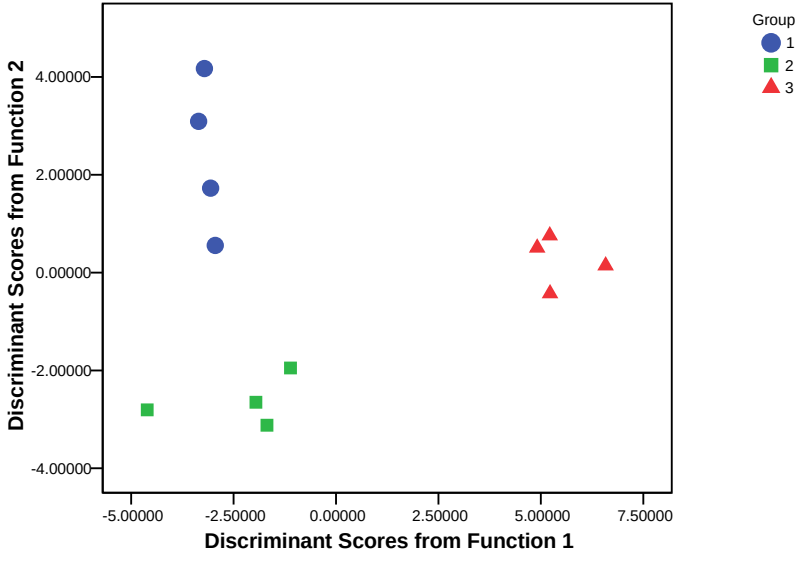
- **NOTE** - if you only have two groups, or if you end up with only **ONE** significant discriminant function – that is, you'd be making a score plot on a line. Boring. Instead, use side-by-side boxplots or other univariate tools to compare groups.

- **NOTE** – if you use quadratic DA, then calculating scores doesn't make sense, since you're projecting onto non-linear space. You can still predict group membership, calculate MH distance, etc, but there isn't a 'linear' direction of discrimination. **SO – no score plots!**

***Example: Iris Data*.** *Only two discriminant functions (and second is probably not significant – note that there is not much discrimination in the direction of the second functions). Because we started with two dimensions, this is basically a rotation.*

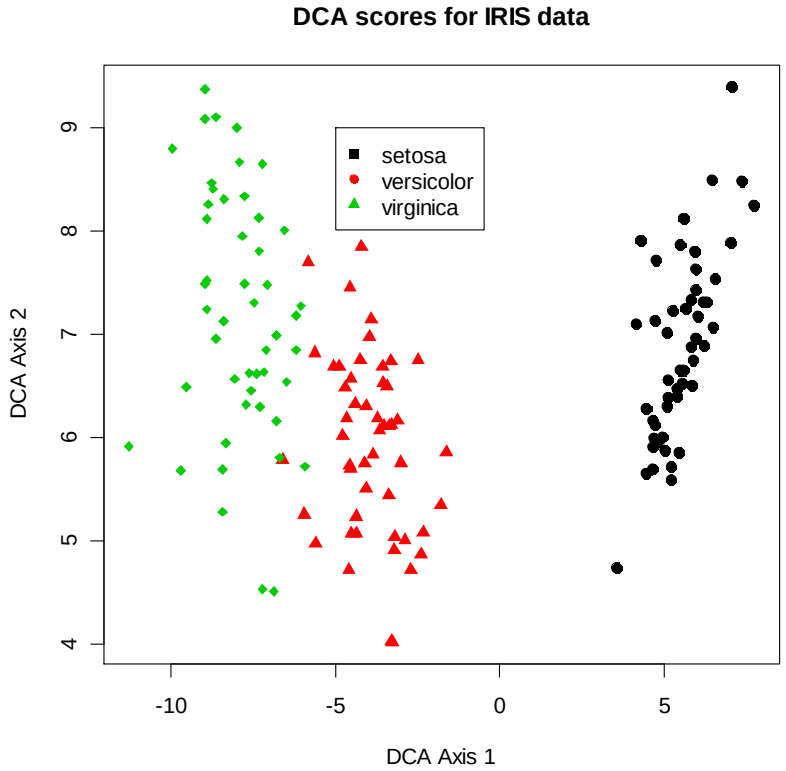


***Example: Heavy Metals*.** *Two discriminant functions, all variables (i.e. we use all metals). Reduces an 8-dimenasional space to a two-dimensional space. Note that this plot did NOT use stepwise DA, which would probably be better – we're using 8 variables and we only have 12 data points (think about over-fitting a regression model . . .)*



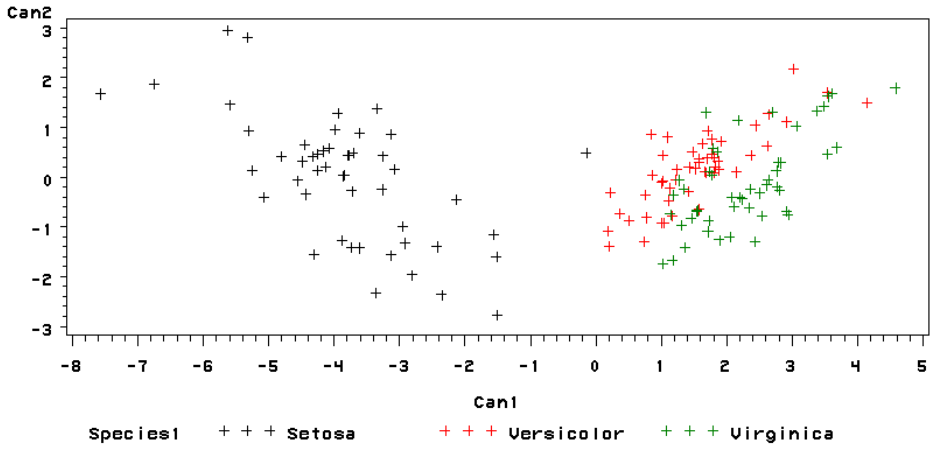
Discriminant Score Plots in R: The `lda()` function will save the coefficients you need to create the scores . These coefficients/weights are called the scaling. Then, plot scores, make graph nice.

NOTE - this does **NOT** work for quadratic discriminant analysis.



Discriminant scores in SAS: In Proc DISCRIM, use the CANONICAL option (this performs linear discriminant analysis) and then use the OUT=*dataset name* option to save the discriminant scores.

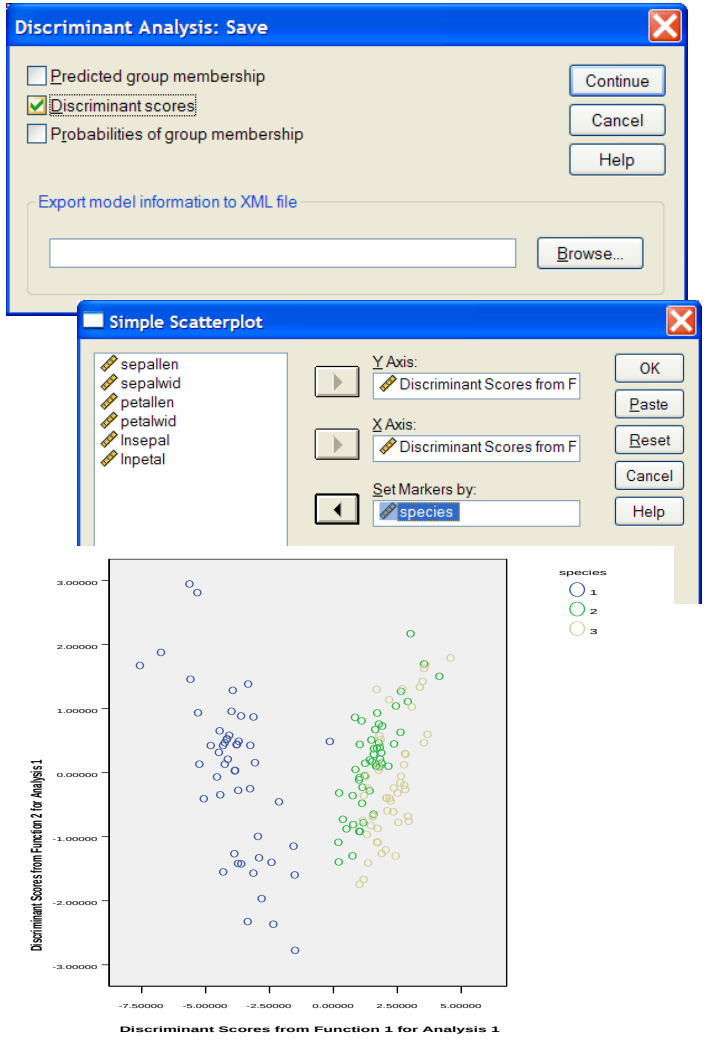
```
PROC DISCRIM DATA=IRIS CANONICAL OUT=OUTIRIS;
  CLASS SPECIES;
  VAR LNSEPAL LNPETAL;
RUN;

PROC GPLOT DATA=OUTIRIS;
  PLOT CAN2*CAN1=SPECIES;
RUN;
```



Discriminant Scores in SPSS: When performing DCA, click on `SAVE` and then click on `Discriminant scores.`

Then use `Graphs` → `Scatterplot` to graph the saved scores.

WARNING – if you perform quadratic DA in SPSS and you ask for scores, it gives you the scores for LINEAR DA (this seems like a flaw to me).



FINAL Comment on Discriminant Analysis

- For two groups, if you have a mixture of categorical and continuous variables, it is often better to use **logistic regression** – this method makes no assumptions about multivariate normality in each group, **AND** allows you to use categorical variables as predictors of group membership.

The End of DA